Cornell University
Cornell Institute for Social and Economic Research

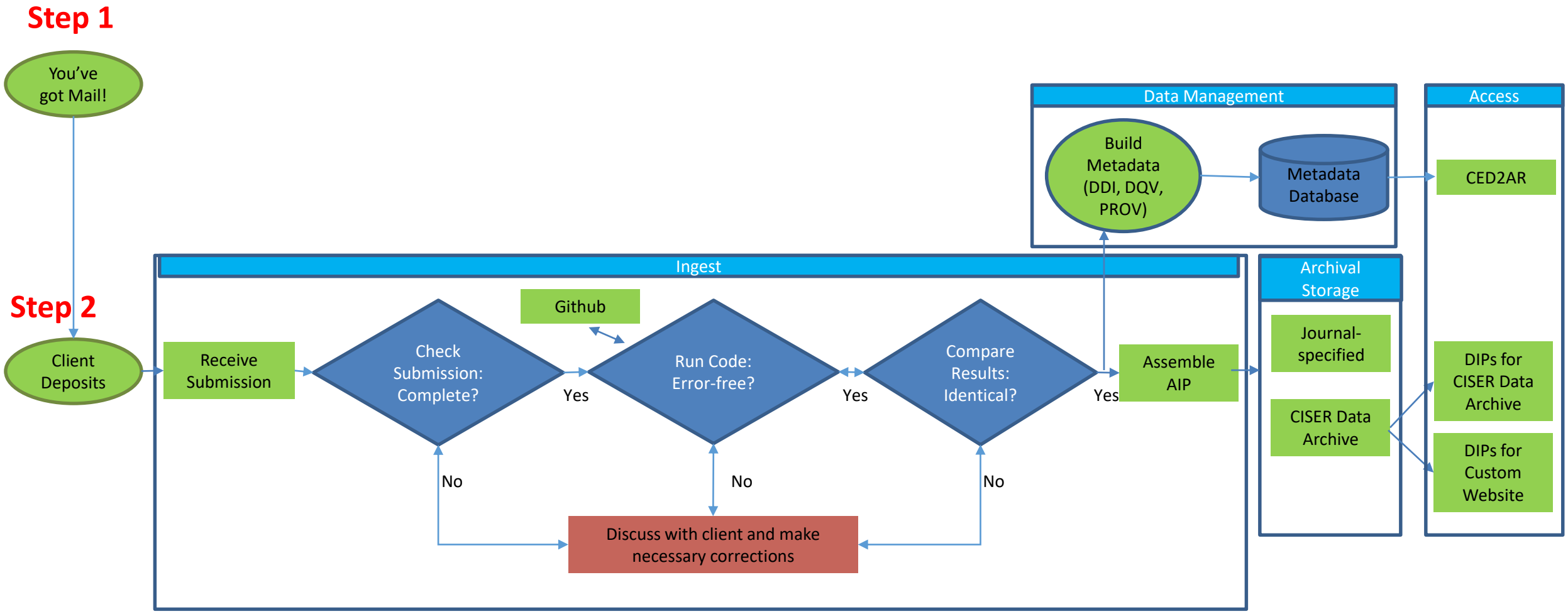
R^2

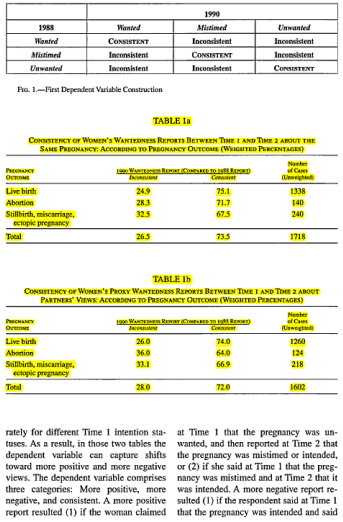
Towards Fully Replicable Data Analysis in an Increasingly Connected World

Florio O. Arguillas, Jr., Ph.D.
William C. Block, Ph.D.
Jeremy Williams
CISER

2016 IASSIST Conference
June 1-3
Scandic Hotel
Bergen, Norway

International Association for Social Science
Information Services & Technology


Step 1
You've got Mail!
Step 2
Client Deposits
Data Management
Build Metadata (DDI, DQV, PROV)
Metadata Database
Access
CED2AR
Ingest
Github
Receive Submission
Check Submission: Complete?
Yes
No
Run Code: Error-free?
Yes
No
Compare Results: Identical?
Yes
No
Assemble AIP
Discuss with client and make necessary corrections
Archival Storage
Journal-specified
CISER Data Archive
DIPs for CISER Data Archive
DIPs for Custom Website



	1990		
1988	*Wanted*	*Mistimed*	*Unwanted*
Wanted	CONSISTENT	Inconsistent	Inconsistent
Mistimed	Inconsistent	CONSISTENT	Inconsistent
Unwanted	Inconsistent	Inconsistent	CONSISTENT

FIG. 1.—First Dependent Variable Construction

TABLE 1a

CONSISTENCY OF WOMEN'S WANTEDNESS REPORTS BETWEEN TIME 1 AND TIME 2 ABOUT THE SAME PREGNANCY: ACCORDING TO PREGNANCY OUTCOME (WEIGHTED PERCENTAGES)

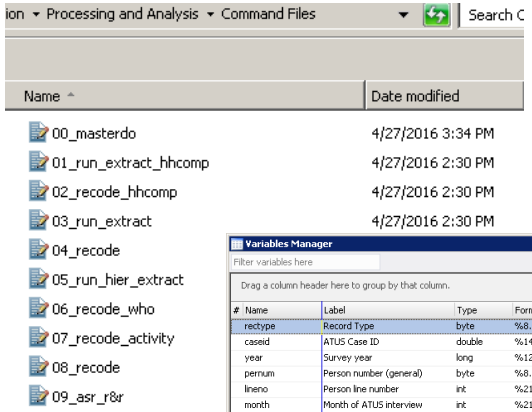
PREGNANCY OUTCOME	1995 WANTEDNESS REPORT (COMPARED TO 1988 REPORT) INCONSISTENT	CONSISTENT	Number of Cases (Unweighted)
Live birth	24.9	75.1	1338
Abortion	28.3	71.7	140
Stillbirth, miscarriage, ectopic pregnancy	32.5	67.5	240
Total	26.5	73.5	1718

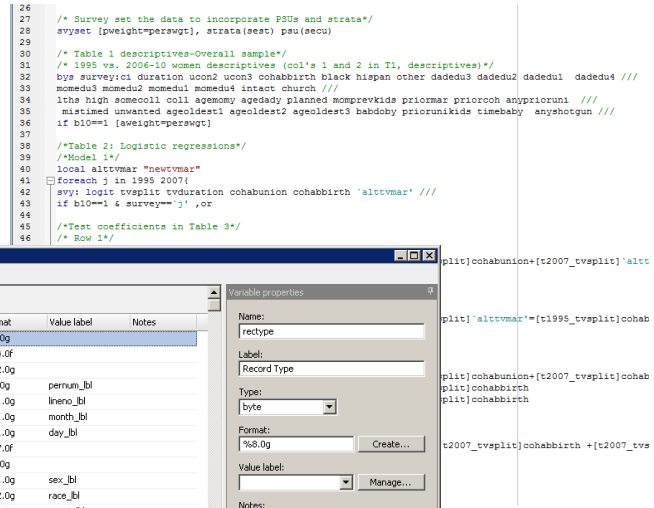
TABLE 1b

CONSISTENCY OF WOMEN'S PROXY WANTEDNESS REPORTS BETWEEN TIME 1 AND TIME 2 ABOUT PARTNERS' VIEWS ACCORDING TO PREGNANCY OUTCOME (WEIGHTED PERCENTAGES)

PREGNANCY OUTCOME	1995 WANTEDNESS REPORT (COMPARED TO 1988 REPORT) INCONSISTENT	CONSISTENT	Number of Cases (Unweighted)
Live birth	26.0	74.0	1260
Abortion	36.0	64.0	124
Stillbirth, miscarriage, ectopic pregnancy	33.1	66.9	218
Total	28.0	72.0	1602

rately for different Time 1 intention statuses. As a result, in those two tables the dependent variable can capture shifts toward more positive and more negative views. The dependent variable comprises three categories: More positive, more negative, and consistent. A more positive report resulted (1) if the woman claimed at Time 1 that the pregnancy was unwanted, and then reported at Time 2 that the pregnancy was mistimed or intended, or (2) if she said at Time 1 that the pregnancy was mistimed and at Time 2 that it was intended. A more negative report resulted (1) if the respondent said at Time 1 that the pregnancy was intended and said



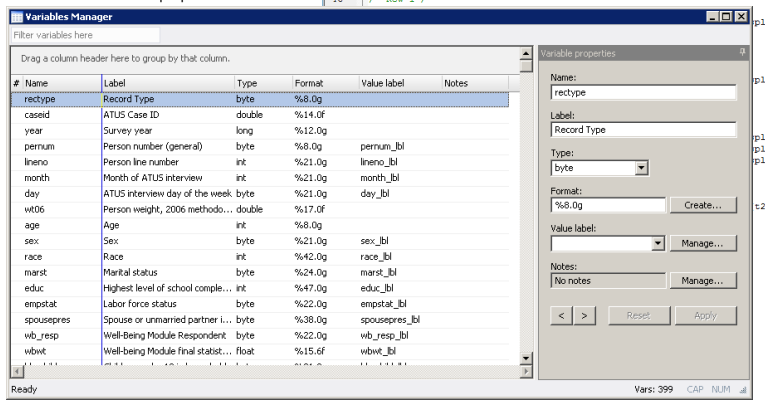


```
/* Survey set the data to incorporate PSUs and strata*/
svyset [pweight=perswgt], strata(sest) psu(secu)

/* Table 1 descriptives-Overall sample*/
/* 1995 vs. 2006-10 women descriptives (col's 1 and 2 in T1, descriptives)*/
bys survey:ci duration ucon2 ucon3 cohabbirth black hispan other dadedu3 dadedu2 dadedu1  dadedu4 ///
momedu3 momedu2 momedu1 momedu4 intact church ///
lths high somecoll coll agemomy agedady planned momprevkids priormar priorcoh anyprioruni  ///
 mistimed unwanted ageoldest1 ageoldest2 ageoldest3 babdoby priorunikids timebaby  anyshotgun ///
if b10==1 [aweight=perswgt]

/*Table 2: Logistic regressions*/
/*Model 1*/
local alttvmar "newtvmar"
foreach j in 1995 2007{
svy: logit tvsplit tvduration cohabunion cohabbirth `alttvmar' ///
if b10==1 & survey=="j" ,or

/*Test coefficients in Table 3*/
/* Row 1*/
```



COMPUTE TOTALINCOME = INCOME1 + INCOME2 + INCOME3.

COMPUTE TOTALINCOME = SUM(INCOME1, INCOME2, INCOME3).

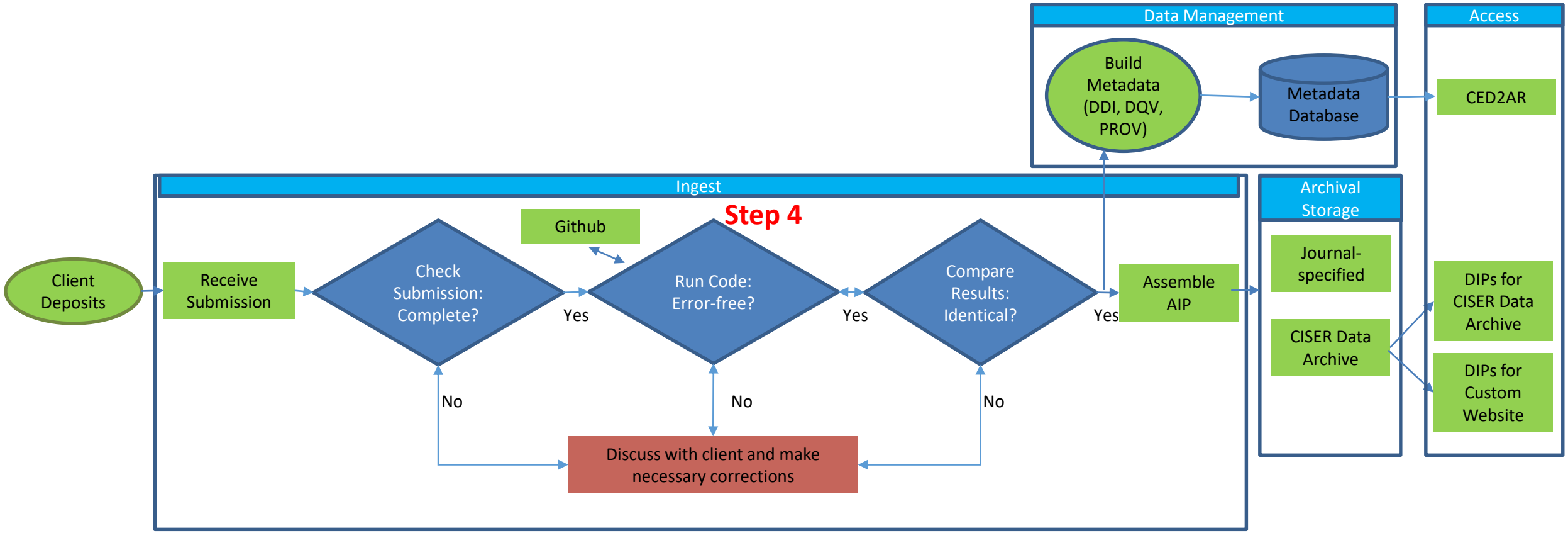


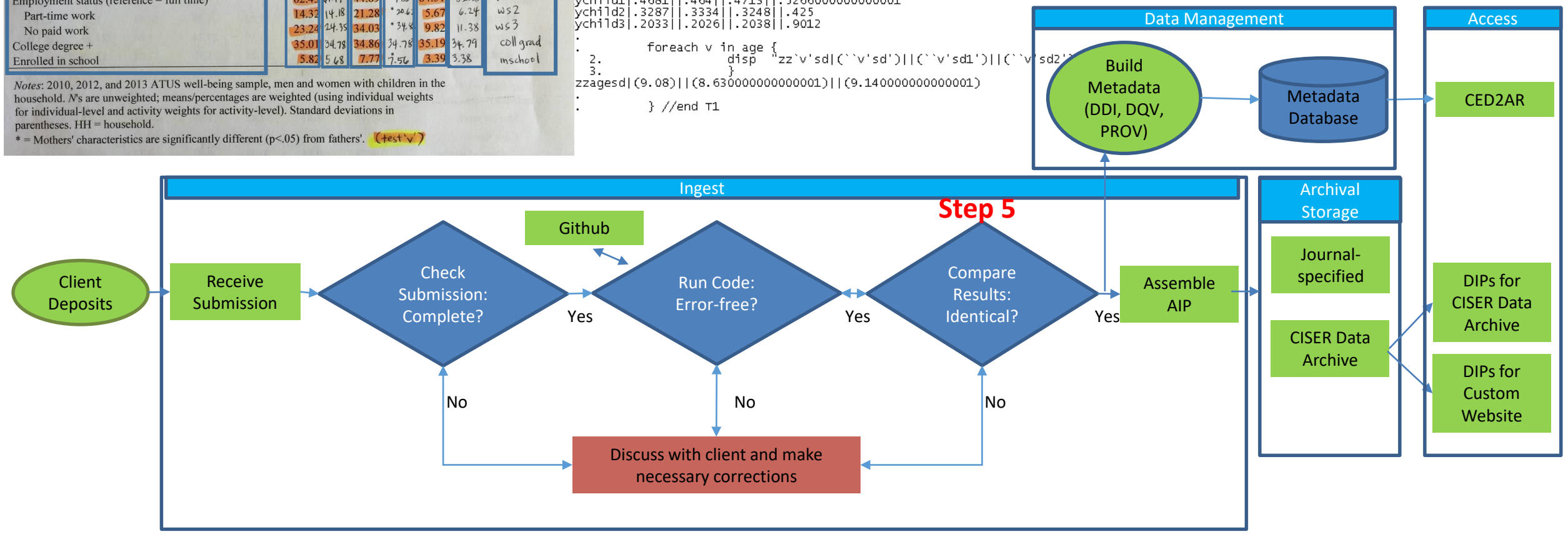
Table 1. Socio-demographic characteristics of parents (*N* individuals = 12,163)

	Full Sample	Mothers	test `v'	Fathers		actP
N (individuals)	12163		7232		4931	
Age	38.38	38.48	36.91	*6.95	40.22	40.35	age
	(9.21)	9.08	9.26)	9.14	(8.76)	8.63	zzagesd
Number of children in the HH (reference = 1 child)	40.05	40.48	41.97	*4.73	37.67	38.95	nkidhh31
2 children	39.50	38.86	39.18	38.51	39.91	39.29	nkidhh32
3+ children	20.44	20.66	18.85	*19.76	22.43	21.76	nkidhh33
Age of youngest child (reference = <6)	47.53	46.81	47.68	47.13	47.33	46.4	ychild1
6-12	32.83	32.87	31.71	32.48	34.22	33.34	ychild2
13-17	19.65	20.33	20.61	20.38	18.44	20.26	ychild3
Race/ethnicity (reference = White, non-Hispanic)	63.44	61.64	62.05	*59.54	65.16	64.25	hr1
Black, non-Hispanic	9.88	10.33	11.72	*12.45	7.60	7.72	hr2
Hispanic	20.11	21.31	20.43	*21.42	19.71	21.17	hr3
Other, non-Hispanic	6.57	6.72	5.80	6.59	7.53	6.88	hr4
Spouse or partner in the HH	82.80	82.6	74.10	*73.99	93.64	93.23	pinhh
Household income (reference = <$25,000)	18.00	18.48	21.32	*22.24	13.87	13.64	inccat1
$25,000-$99,999	57.70	56.2	56.32	*54.39	59.42	58.67	inccat2
$100,000 +	23.14	24.2	20.78	*21.02	26.08	26.91	inccat3
Income missing	1.16	1.02	1.58	*1.19	0.64	0.82	inccat4
Employment status (reference = full time)	62.43	61.47	44.69	*44.5	84.51	82.38	ws1
Part-time work	14.32	14.18	21.28	*20.62	5.67	6.24	ws2
No paid work	23.24	24.35	34.03	*34.8	9.82	11.38	ws3
College degree +	35.01	34.78	34.86	34.78	35.19	34.79	collgrad
Enrolled in school	5.82	5.68	7.77	*7.56	3.39	3.38	mschool

Notes: 2010, 2012, and 2013 ATUS well-being sample, men and women with children in the household. *N*s are unweighted; means/percentages are weighted (using individual weights for individual-level and activity weights for activity-level). Standard deviations in parentheses. HH = household.

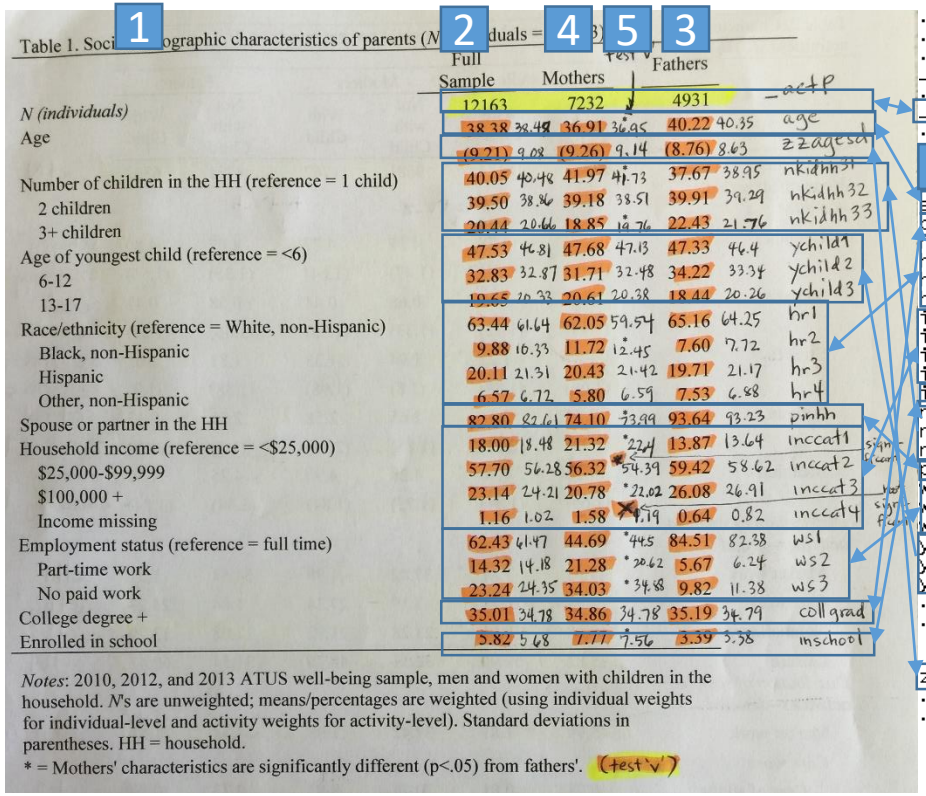
* = Mothers' characteristics are significantly different (p<.05) from fathers'. (test `v')

```
.         **print Table 1
.         disp  "__varname|fs||men||women||diff m v. w"
__varname|fs||men||women||diff m v. w
.         disp "_actP|`Nt'||`Nt1'||`Nt2'"
_actP|12163||4931||7232
.
.         foreach v in `a1vars' {
  2.                 disp  "`v'|``v'f'||``v'1'||``v'2'||`test`v''"
  3.                 }
age|38.4752||40.3477||36.9548||0
collgrad|.3478||.3479||.3478||.9938
hr1|.6164000000000001||.6424000000000001||.5954||.0001
hr2|.1033||.0772||.1245||0
hr3|.2131||.2117||.2142||.8134
hr4|.0672||.0688||.0659||.6115
inccat1|.1848||.1364||.224||0
inccat2|.5628000000000001||.5862000000000001||.5439000000000001||.0002
inccat3|.2421||.2691||.2202||0
inccat4|.0102||.0082||.0119||.1349
inschool|.0568||.0338||.0756||0
nkidhh31|.4048||.3895||.4173||.0152
nkidhh32|.3886||.3929||.3851||.4867
nkidhh33|.2066||.2176||.1976||.0379
pinhh|.8261000000000001||.9323||.7399||0
ws1|.6147||.8238000000000001||.445||0
ws2|.1418||.0624||.2062||0
ws3|.2435||.1138||.3488||0
ychild1|.4681||.464||.4713||.5266000000000001
ychild2|.3287||.3334||.3248||.425
ychild3|.2033||.2026||.2038||.9012
.
.         foreach v in age {
  2.                 disp  "zz`v'sd|(``v'sd')||(``v'sd1')||(``v'sd2')"
  3.                 }
zzagesd|(9.08)||(8.630000000000001)||(9.140000000000001)
.
.         } //end T1
```



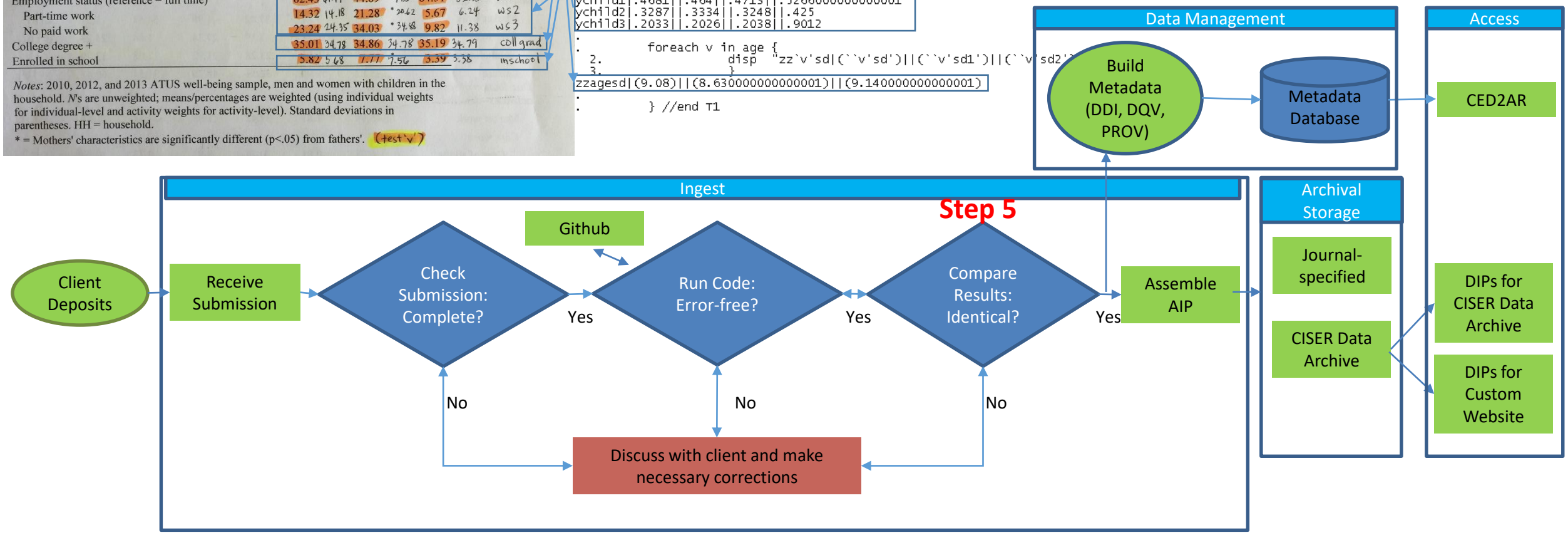
iassist

International Association for Social Science Information Services & Technology



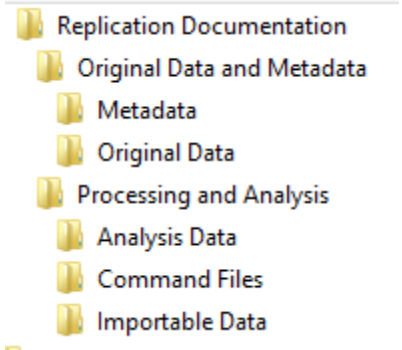

```
.         **print Table 1
.         disp  "__varname|fs||men||women||diff m v. w"
__varname|fs||men||women||diff m v. w
.         disp "_actP|`Nt'||`Nt1'||`Nt2'"
_actP|12163||4931||7232
.
.         foreach v in `varlist' {
.                 disp "`v'|"``v'f'||``v'1'||``v'2'||`test`v''"
.                 }
age|38.4752||40.3477||36.9548||0
collgrad|.3478||.3479||.3478||.9938
hr1|.6164000000000001||.6424000000000001||.5954||.0001
hr2|.1033||.0772||.1245||0
hr3|.2131||.2117||.2142||.8134
hr4|.0672||.0688||.0659||.6115
inccat1|.1848||.1364||.224||0
inccat2|.5628000000000001||.5862000000000001||.5439000000000001||.0002
inccat3|.2421||.2691||.2202||0
inccat4|.0102||.0082||.0119||.1349
inschool|.0568||.0338||.0756||0
nkidhh31|.4048||.3895||.4173||.0152
nkidhh32|.3886||.3929||.3851||.4867
nkidhh33|.2066||.2176||.1976||.0379
pinhh|.8261000000000001||.9323||.7399||0
ws1|.6147||.8238000000000001||.445||0
ws2|.1418||.0624||.2062||0
ws3|.2435||.1138||.3488||0
ychild1|.4681||.464||.4713||.5266000000000001
ychild2|.3287||.3334||.3248||.425
ychild3|.2033||.2026||.2038||.9012
.
.         foreach v in age {
  2.                 disp  "zz`v'sd|(``v'sd')||(``v'sd1')||(``v'sd2')"
  3.                 }
zzagesd|(9.08)||(8.630000000000001)||(9.140000000000001)
.
.         } //end T1
```

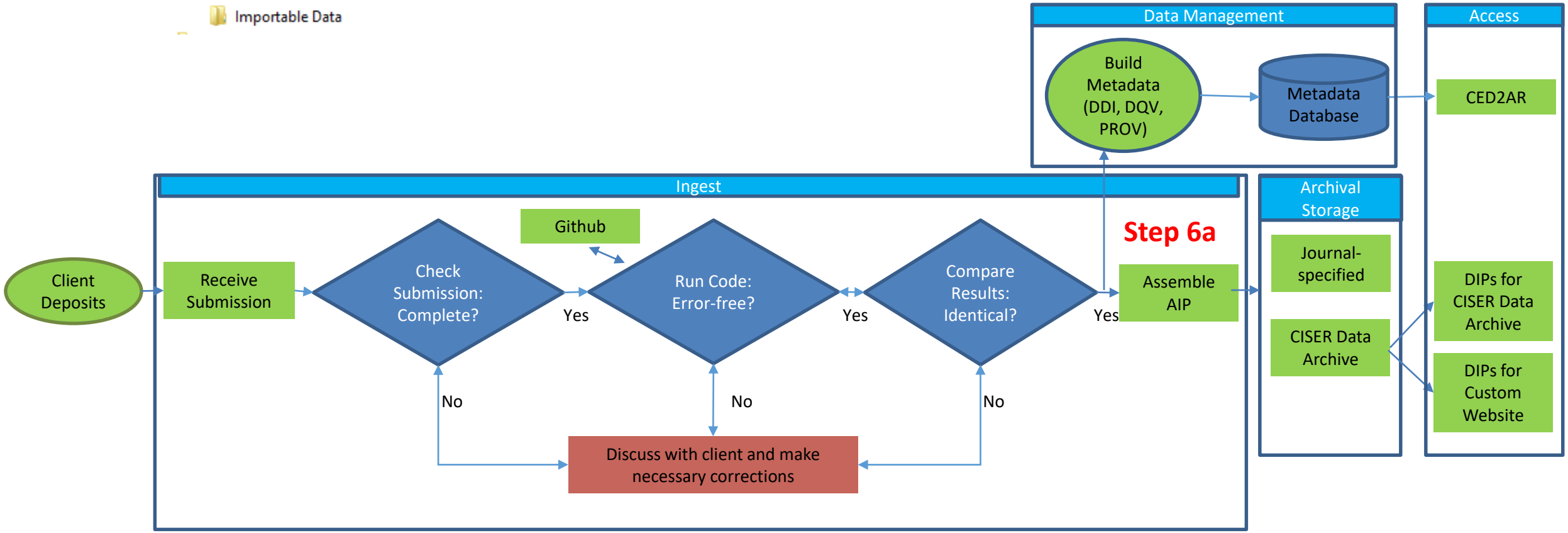
International Association for Social Science Information Services & Technology



Teaching Integrity in Empirical Research (TIER) protocol

https://www.haverford.edu/project-tier/protocol-v2

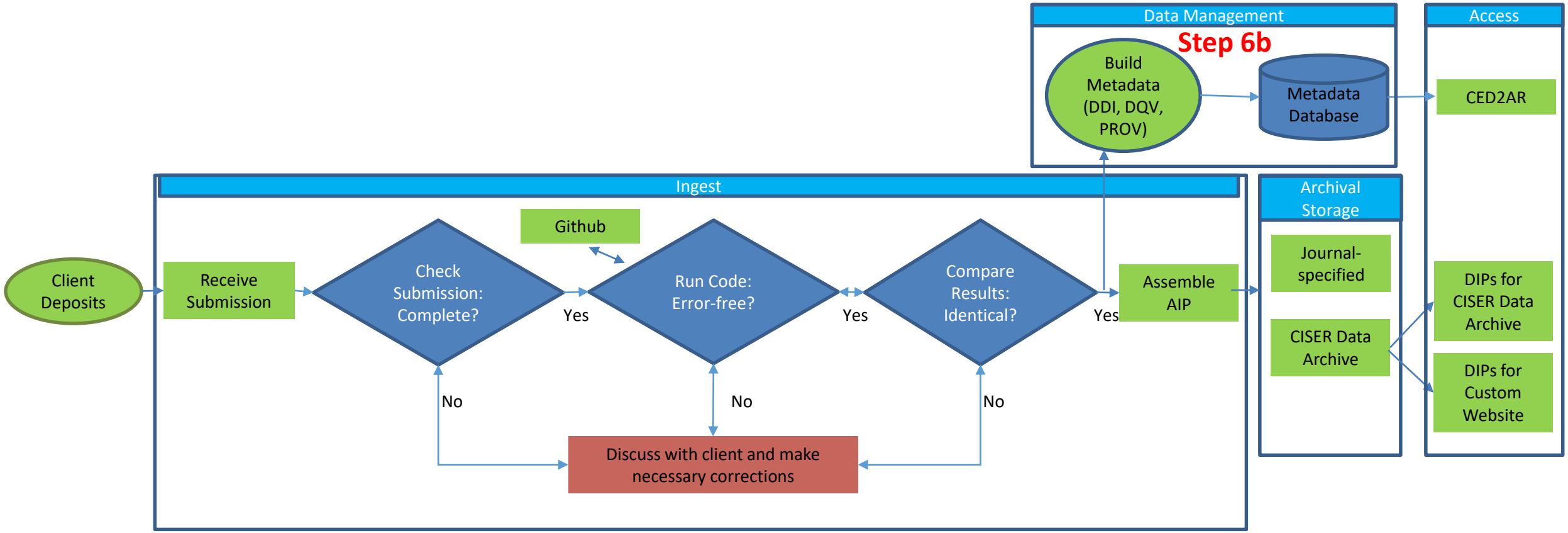


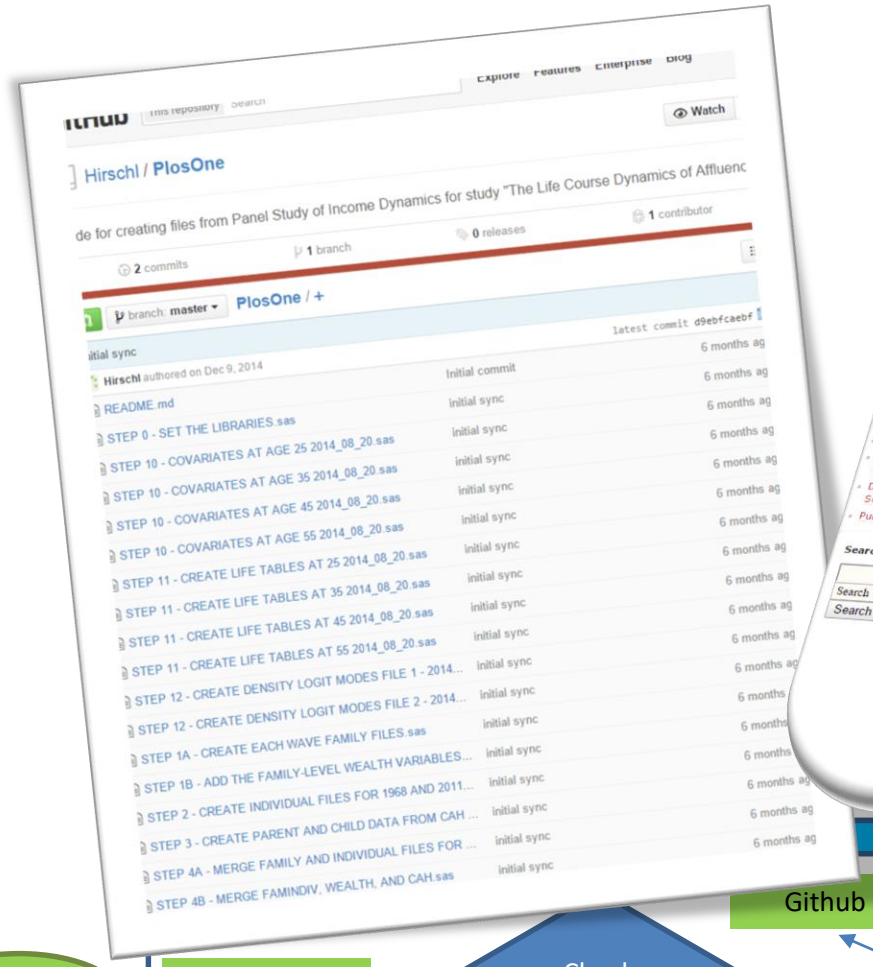


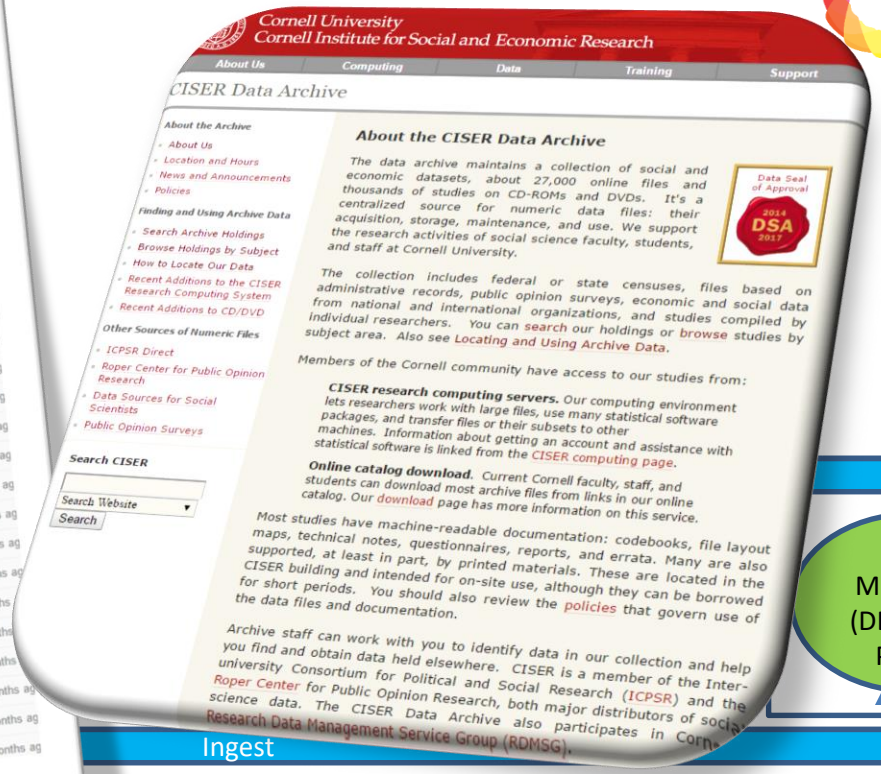
Comprehensive Extensible Data Documentation and Access Repository (CED^{2}AR)

Data Quality Vocabulary (DQV): http://w3c.github.io/dwbp/vocab-dqg.html

PROV: http://www.w3.org/TR/prov-primer/







International Association for Social Science Information Services & Technology

Data Management

Build Metadata (DDI, DQV, PROV)

Metadata Database

Access

CED2AR

Ingest

Github

Client Deposits

Receive Submission

Check Submission: Complete?

Yes

Run Code: Error-free?

Yes

Compare Results: Identical?

Yes

Assemble AIP

No

No

No

Discuss with client and make necessary corrections

Archival Storage

Journal-specified

Step 7

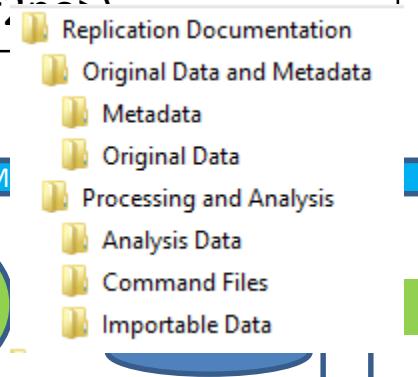
CISER Data Archive

DIPs for CISER Data Archive

DIPs for Custom Website

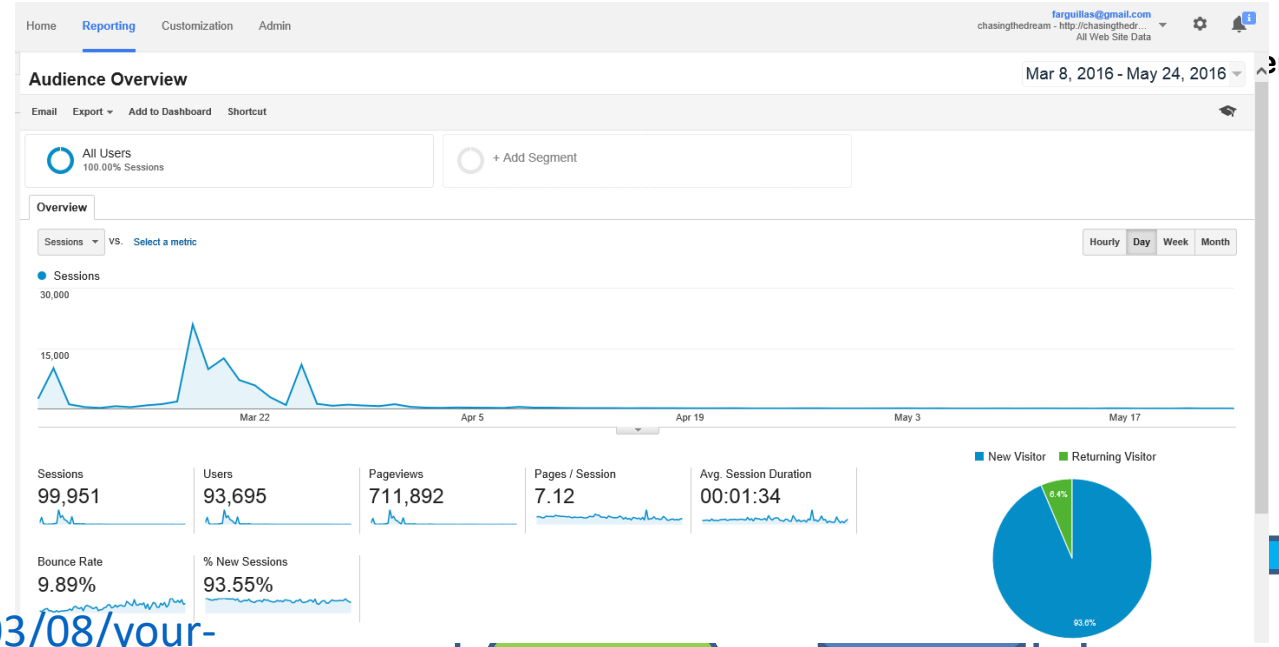
Links to:

a) Metadata with variable level search through CISER's instance of CED2AR
b) Journal article
c) Replication materials: V:\R2\<1stAuth>\<r2no>\



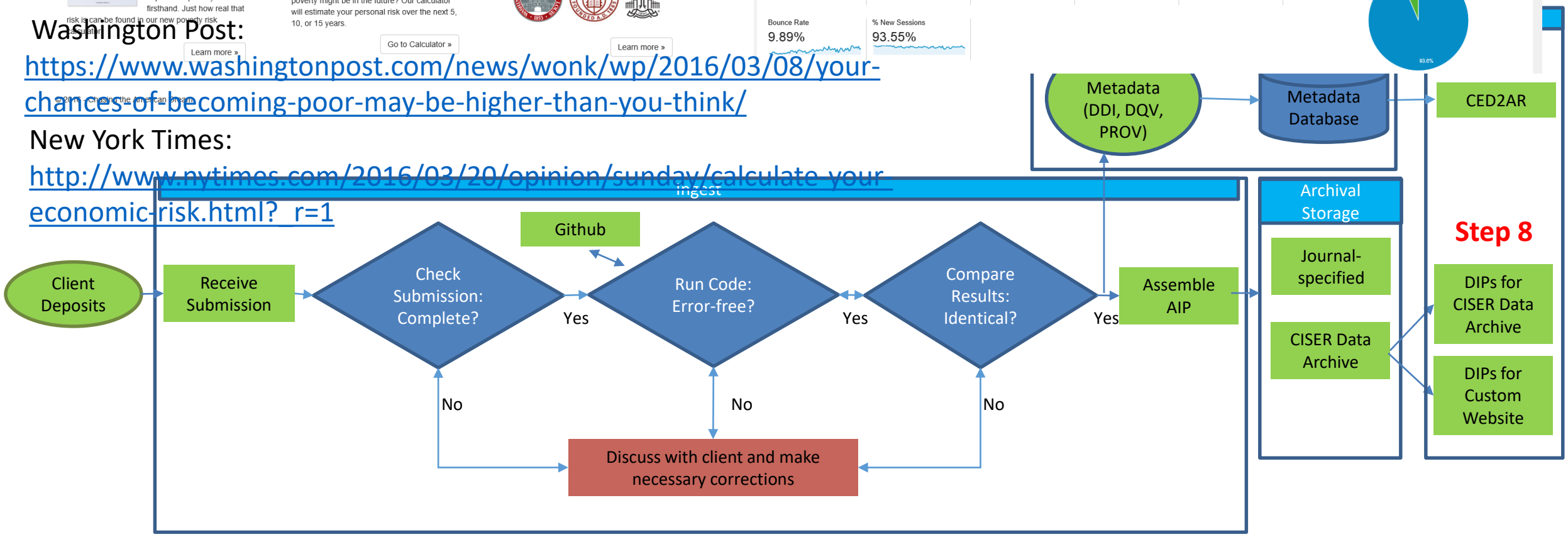
Will also include:

- Persistent URL: http://doi.org/10.6077/J5/CISER2775
- Study citations
- Data citation
- Documentation about known quality issues
- User Feedback Form
- Forum or mailing list
- Social Media Accounts (e.g., Facebook, Twitter)



Washington Post:
https://www.washingtonpost.com/news/wonk/wp/2016/03/08/your-chances-of-becoming-poor-may-be-higher-than-you-think/

New York Times:
http://www.nytimes.com/2016/03/20/opinion/sunday/calculate-your-economic-risk.html?_r=1



Conclusion

- Process is time-intensive, costly, and confusing if data and code quality is low
- Suggest code review, use of versioning software, adherence to TIER protocol in managing files
- Reach out to Cornell graduate students and provide training on how to prepare research materials for reuse thru classes or workshops.
- CURATE AS YOU CODE AND CODE WITH REUSE IN MIND.
- For archives:
  - Incorporate DQV and PROV
  - Follow TOP Guidelines, Open Data Certificate Requirements, and TIER Protocl for managing and packaging replication materials
  - Anything else?

Thank you!

- Contact author: Florio Arguillas (foa2@cornell.edu)